# Act Report

## Introduction and Background

The aim of this project is to wrangle data, that is, gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, and clean it using Python and its libraries. The project also includes data analyses and visualizations.

The dataset used is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators can have very different values, even larger than 10. This is part of the account's humorous style.

## Gathering, Assessing and Cleaning the Data

For a description of the wrangling process see that file 'wrangle_report.'
To see all the details of the process see the file 'wrangle_act.'

## Analysis and Visualizations

The cleaned data was stored in a unique data set and was ready for analysis.
This data set includes 2084 unique rows and 12 columns.

### *Rating Numerator, Retweets and Favorite Counts*

| | rating_numerator | retweet_count | favorite_count |
|---|---|---|---|
| **count** | 2084.000000 | 2084.000000 | 2084.000000 |
| **mean** | 10.605086 | 2502.307102 | 8287.873800 |
| **std** | 2.148360 | 4388.090377 | 12060.459635 |
| **min** | 0.000000 | 11.000000 | 70.000000 |
| **25%** | 10.000000 | 556.750000 | 1837.250000 |
| **50%** | 11.000000 | 1214.000000 | 3777.500000 |
| **75%** | 12.000000 | 2844.250000 | 10365.500000 |
| **max** | 14.000000 | 77210.000000 | 155268.000000 |

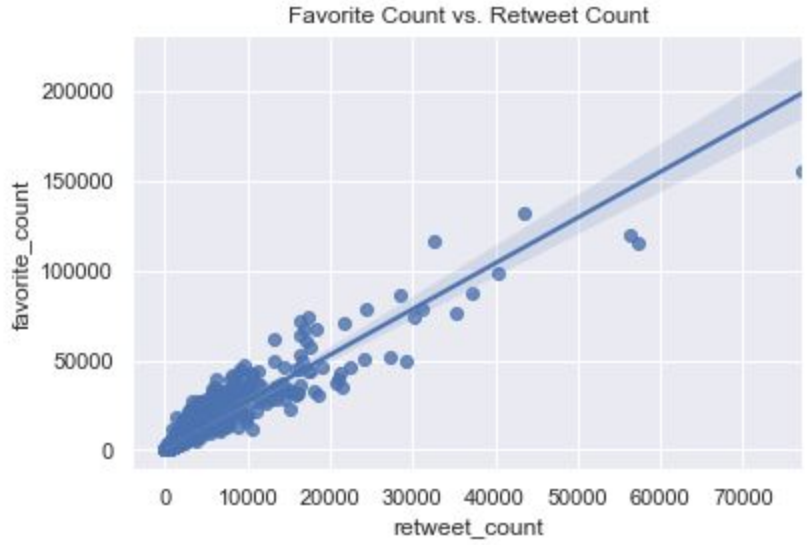
From the summary above we can already gain several insights into the data:

1. The mean rating numerator is 10.6 and the standard deviation, which means that we can most likely expect to find values from 8 to 12. The lowest value is 0 and the highest 14. We can see from the percentiles that values seem to be concentrated around 10, 11 and 12.

2. The count of retweets is on average 2502.3, with a standard deviation of 4388.1, which is high. We can see that there is a greater difference between the minimum value and the maximum value.

3. The count of favorites is quite larger than retweets. This could point to the fact that people are more likely to mark a tweet as a favorite than to retweet them. As with the count of retweets, the standard deviation for favorite counts is very large, and the minimums and maximums are very far apart.

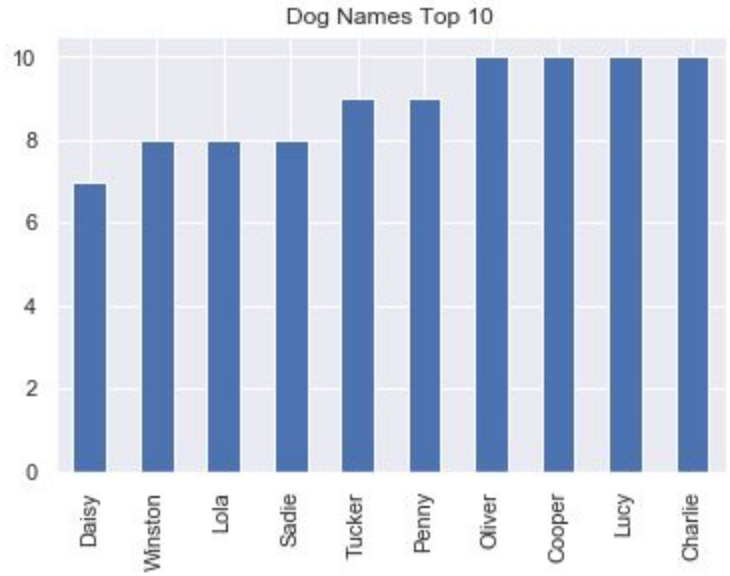
***Correlation Between Retweets and Favorites***




The scatter plot above shows a clear strong positive correlation between the count of favorites and of retweets. This means that, in general, the higher the number of favorite counts that higher the number of retweets and vice versa.

***Dog Names Top 10***

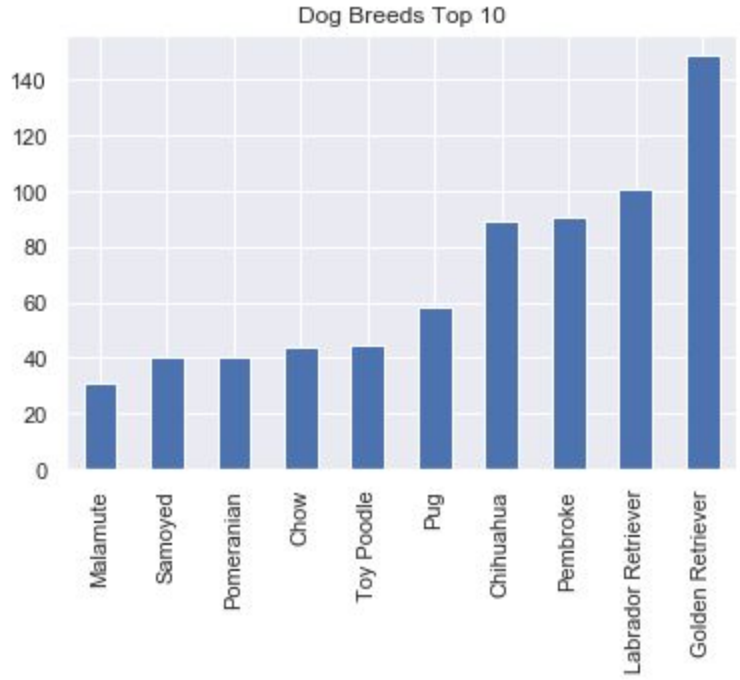
The top 10 of the most popular dog names in the data set looks as follows:




The four most popular names, with 10 appearances each, are: Charlie, Lucy, Cooper and Oliver.

***Dog Breeds Top 10***

The top 10 of the most popular dog breeds in the data set looks as follows:




The Golden Retriever is clearly the most popular dog breed, with a difference of more than 40 appearances than the second, the Labrador Retriever.